Data Integration Primer

U.S. Department
of Transportation

Data Integration Primer

OFFICE OF
ASSET
MANAGEMENT

U.S. Department
of Transportation

August 2001

U.S. Department of Transportation
Federal Highway Administration
Office of Asset Management

TABLE OF CONTENTS

NOTE FROM THE DIRECTOR

Office of Asset Management, Infrastructure Core Business Unit, Federal Highway Administration

Data integration is a fundamental component of Transportation Asset Management. A common and consolidated set of data from which to draw information enhances a transportation agency's ability to make well-informed, cost-effective Asset Management decisions. Improvements in computer data processing and information management technologies have made it possible for organizations to electronically link or combine large data files. For this reason the Federal Highway Administration's (FHWA) Office of Asset Management has prepared this Primer on data integration to help State and local transportation agencies understand the importance of integrated databases and to provide options for developing or expanding existing data integration initiatives.

Currently, little literature is available that addresses the many facets of database integration, particularly in relation to Asset Management. This Primer answers such key questions as, "How can data integration improve the Asset Management process in my agency?" "What important steps do we need to follow?" "What hardware, tools, and software options are available?" "What potential obstacles can hinder the process?" and "How can we overcome such roadblocks?" The Primer also cites recent experiences of State departments of transportation that have integrated some or all of their transportation data. I believe that this Primer will be helpful to administrators and managers interested in data integration benefits, approaches, and impediments.

A separate glossary of common terms related to data integration is provided as a supplement to this Primer. The glossary will assist those who are not familiar with the technical words and phrases used in data integration and information management, some of which appear in the Primer.

The principles, options, and barriers described in this Primer highlight many issues related to specific components of data integration that will require in-depth investigation. Such issues include location referencing, database standards, staffing, and other technical and institutional considerations. FHWA's Office of Asset Management will provide assistance to transportation agencies in addressing these and other data integration issues. We are prepared to work with our partners to promote the use of best-practice data integration in Asset Management.

Madeleine Bloom

Madeleine Bloom
Director, Office of Asset Management

OVERVIEW

WHAT IS ASSET MANAGEMENT?

Asset Management is a framework for making cost-effective resource allocation, programming, and management decisions. It combines engineering principles with sound business practices and economic theory, and provides tools to facilitate a more organized, logical, and comprehensive approach to decision-making.

Managing transportation assets is not a new idea. Highway agencies have been developing and utilizing pavement, bridge, and maintenance management systems for at least the past two decades. Asset Management—that is with a capital "A" and a capital "M"—implies the merging of these individual asset management systems into a unified management approach.

The Asset Management process is represented by the flowchart in Figure 1, Asset Management Framework. The flowchart shows not only the individual Asset Management components or business practices but also their relationships and the order in which they will likely occur. The goals and policies of the agency consist of high-level, strategic statements that reflect the desired condition and performance of the transportation system from the perspectives of both the agency and its customers. These goals and policies therefore guide how the assets are managed at all levels of the organization.

An inventory of the assets and a means to assess their condition and model their performance enable the agency to identify investment requirements for improvement in the short and long term.

Next, the agency identifies the options or alternatives for addressing the investment requirements, which are analyzed and evaluated on the basis of their cost-effectiveness using a host of analytical and optimization tools. Budget and resource allocation constraints are incorporated into the alternatives evaluation criteria. Selected alternatives are included in the list of projects that will go into the agency's short- and long-range plans. The final stages in the process consist of implementing the projects and monitoring the resulting performance of the assets.

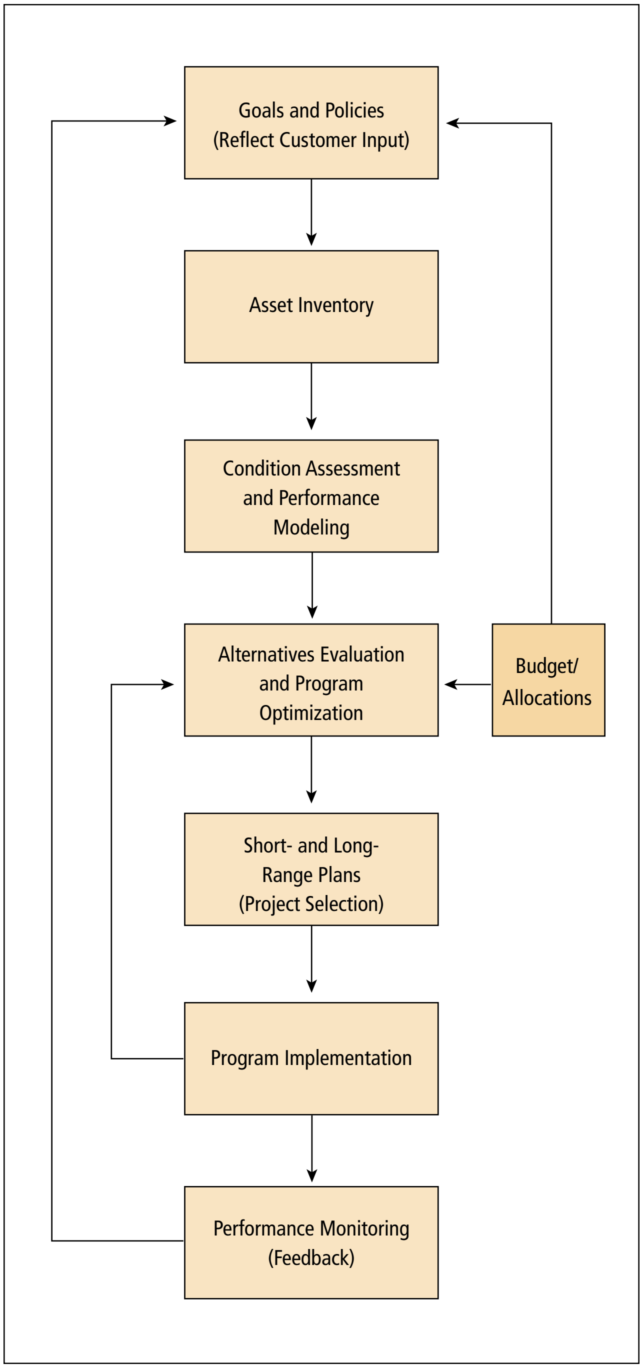

Goals and Policies (Reflect Customer Input)
Asset Inventory
Condition Assessment and Performance Modeling
Alternatives Evaluation and Program Optimization
Budget/ Allocations
Short- and Long-Range Plans (Project Selection)
Program Implementation
Performance Monitoring (Feedback)

FIGURE 1. ASSET MANAGEMENT FRAMEWORK

The knowledge gained from one cycle in the process is used to update and improve any or all of its components. Figure 1 depicts a generic Asset Management process in a transportation agency. However, the manner in which each component is carried out will vary from one agency to another.

WHY IS ASSET MANAGEMENT IMPORTANT?

The time is right for Asset Management in transportation agencies. Demands on the nation's mature and deteriorating transportation system have increased considerably due to high and growing transport productivity and mobility. These increased demands present a significant challenge to many agencies, especially where budget and personnel constraints force them to do more asset enhancement and preservation work with less staff and less money. Asset Management helps agencies identify ways to leverage their resources to respond to increasing system demands while maintaining adequate levels of service. It provides a means to prioritize requirements and allocate funds across different assets and over time in the most cost-effective way.

Asset Management also allows agencies to take advantage of increasingly powerful and generally affordable computers, sophisticated analytical tools, and advances in information technology. The new technology provides quicker and improved ways to gather, process, and analyze data as well as to make sound management decisions affecting the assets.

Finally, Asset Management helps agencies demonstrate to the public that they are responsible stewards of the Nation's transportation assets. Increasingly, the public is demanding more information about how effectively the government is managing the Nation's infrastructure. Asset Management, which is based on combined engineering and economic principles, will assist in demonstrating to the public that the government is making sound transportation investment decisions.

THE ROLE OF DATA IN ASSET MANAGEMENT

Useful and reliable data are central to a fully functioning Asset Management process. Asset Management is a data-intensive process that involves the gathering, retrieval, storage, analysis, and communication of enormous quantities of data. The information that is drawn from these data is essential to the cooperative and informed decision-making process underlying Asset Management. Information is required to evaluate and monitor the condition and performance of the asset inventory, develop performance objectives and measures, identify cost-effective investment strategies, and conduct asset value assessments. Information is also required to monitor the effectiveness of the Asset Management business process. Although it is not necessary to store all the transportation system's data in a single repository, it is critical that the data be readily accessible and comparable. Data integration and data sharing, therefore, are vital components of Asset Management.

WHAT IS DATA INTEGRATION?

Data integration is the process of combining or linking two or more data sets from different sources to facilitate data sharing, promote effective data gathering and analysis, and support overall information management activities in an organization.

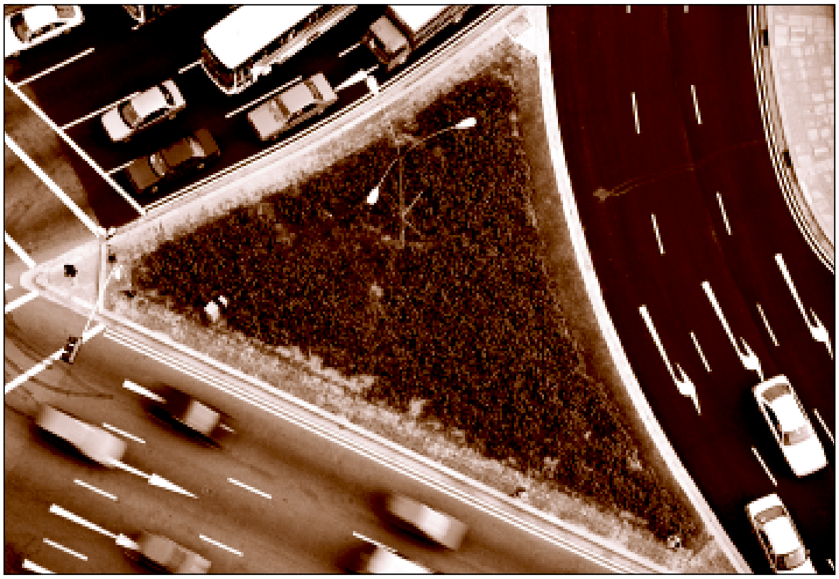
In general, transportation agencies acquire, store, and analyze large amounts of data to support their day-to-day operations as well as to create long-term plans and programs. Each unit within the agency keeps a record of its own data in paper or electronic format. Typically these units exchange and share data among themselves. Making data available and accessible to all the users of data in the organization, however, is required for Asset Management. For example, the agency's maintenance units need information about the physical and design characteristics of maintainable assets such as pavements, bridges, guardrails, and signs. Inventories of these assets may exist separately in the agency's pavement, bridge, and traffic operations units' databases.

Information about the assets and a process for sharing this information are critical to successful Asset Management. The extent to which this data is shared and used to make coordinated decisions determines the effectiveness of the Asset Management process.

Beyond Asset Management, the incentives for data integration are readily apparent to organizations that collect, store, and manage disparate databases. Agencies that combine or link their multiple databases can reduce data collection and management costs, improve the accuracy and timeliness of the information, and support a variety of applications that draw data from various sources.

THE PRIMER

This Primer provides information to all transportation agencies on the principles of data integration and data sharing and their role in comprehensive Asset Management. It supplements FHWA's "Asset Management Primer," which identifies data integration as essential in supporting an enhanced and highly integrated decision-making framework. This work describes the benefits of data integration from a general information management perspective as well as for specific elements of Asset Management. It discusses the key principles and elements of the data integration process and presents the options that are available for implementation. It also describes the most common data integration challenges. Actual State department of transportation (DOT) experiences are incorporated in the discussions to illustrate success stories and lessons learned.



WHY INTEGRATE DATA?

Asset Management relies heavily on highly organized and integrated databases to drive its many decision-support functions. Information systems used for Asset Management, including those for pavements, bridges, tunnels, hardware, and highway maintenance, commonly draw inputs from several data sources in an agency. Many transportation agencies have established databases and collection procedures that support existing management systems and have made significant strides in deploying these systems. However, much remains to be done in terms of establishing mechanisms for bringing the data from these disparate systems to a common decision-making framework. Clearly these individual management systems will benefit from an integrated database environment.

The extent to which an agency may benefit from data integration depends upon the flow of data within the organization; how and how often the data are acquired, processed, stored, and used; and who uses the data. Although there may be disincentives to combining and linking databases, including cost and other burdens integration imposes on the agency, these disincentives are generally far outweighed by the long-term benefits. Below is a brief description of potential benefits—general benefits as well as incentives specific to Asset Management—that could encourage transportation agencies to integrate their databases.

GENERAL DATA INTEGRATION BENEFITS

Effective data integration and data sharing typically result in improved information processing and decision-making capabilities to support Asset Management in an agency. Compared with autonomous databases, integrated databases offer many advantages:

Availability/Accessibility—Because the asset data are easily retrieved, viewed, queried, and analyzed by anyone within the agency, they are in a more user-friendly form. Making data available and accessible opens information up to a larger user community.

The Transportation Management Information System (TMIS) being developed by the Mississippi DOT is an integrated database system that tracks information about the agency's transportation components. TMIS allows users to combine bridge, pavement, and traffic data in one query and have an answer moments later.

Timeliness—Data can be quickly updated and therefore are generally current. Normally in a combined or linked database only one update is necessary. In addition, databases are time-stamped to keep track of when they were created, modified, or updated.

Accuracy, Correctness, and Integrity—Data are commonly free of errors because the integrity of the databases is upheld in an integrated environment. Likewise, processes exist for automatic or convenient error-checking and verification to help maintain the quality of the data during data entry and access.

Consistency and Clarity—Specified data have a clear and unique definition throughout the agency, thereby avoiding or eliminating confusing and conflicting meaning and usage of data.

Completeness—All available information associated with the assets, including historical and recently collected data, can be found in the database. In addition, missing records or data fields are easily identified and flagged by the data management process.

Reduced Duplication—Identical data are not stored in multiple, disparate locations. Integration reduces the need for multiple updating and ensures that everyone has access to the same data.

Faster Processing and Turnaround Time—Less time is spent on consolidating and transmitting data to various users in the agency. The integrated data environment also supports faster data manipulation, allowing multiple users to conduct separate analyses concurrently.

Lower Data Acquisition and Storage Cost—Data are not collected or processed twice, and are consolidated and stored at locations in the agency that provide optimal convenience and ease of maintenance.

Informed and Defensible Decisions—Highly organized, comprehensive databases allow users to drill down through successive levels of detail for any given asset. This depth provides more information to support recommendations and allows users to conduct different types of analysis using various subsets or combinations of data.

Integrated Decision-Making—One of the greatest benefits of data integration is that it permits decision-support analysis throughout the transportation system, from the field to the executive level (vertically) and across divisions within and outside the organization (horizontally).

TRIMS, the Tennessee DOT's Roadway Information Management System, gives everyone in the agency the same asset information base to work from. Any data collected can be input directly from the desktops, updating the system immediately, and everyone is able to access the most current data at any time.

DATA INTEGRATION BENEFITS SPECIFIC TO ASSET MANAGEMENT

Table 1 lists specific Asset Management business processes and how each could potentially be enhanced by data integration based on the general incentives described above. The processes listed are carried out at different levels and by a broad range of staff in an organization. Hence it is apparent that data integration can potentially benefit everyone involved in Asset Management.

One can identify other more specific Asset Management functions that are not shown in Table 1. The important message being conveyed is that agencies interested in implementing successful Asset Management can have a considerably improved process through data integration.

TABLE 1.
BENEFITS OF DATA INTEGRATION TO THE ASSET MANAGEMENT PROCESSES

Asset Management Business Process	Potential Benefits of Data Integration
Inventory and Field Data Collection	Promotes one-time (single source) data acquisition and uploading Allows updating and processing of inventory records using a single transaction Encourages conformity to database quality standards Allows determination of how much data exists and how much needs to be collected Reduces data handling and processing times with built-in data checking and verification Helps standardize the inventory and data collection procedures throughout the agency
Condition Assessment and Investment Requirement Determination	Supports convenient historical and spatial condition analysis Enables quick identification of assets that need immediate attention Allows a more thorough and detailed assessment of investment requirements and minimizes the risk of making incorrect condition assessments Promotes standard condition rating procedures and uniform criteria for evaluation Provides convenient database storage of condition/investment analysis data and results Promotes collective decision-making from various parts of the organization
Identification and Selection of Strategies	Allows combining data about previous activities or decisions made on the assets with existing condition data to develop more effective management strategies Reduces the risk of choosing inappropriate or ineffective actions Prevents the inadvertent development of multiple strategies for the same asset Facilitates economic evaluation of alternatives Permits results of analyses to be readily stored and retrieved Provides for fact-based strategies
Program Development	Improves statewide program development through comprehensive system information Provides timely information for high-priority programs and plans Promotes efficient distribution of funding among competing programs Promotes consistency in programs from year to year and across organizational units Offers a convenient repository for past and current transportation programs Helps support choice of transportation programs Allows evaluation of multiple program categories
Scheduling of Activities and Allocation of Resources	Expedites timing of activities and assignment of resources Helps identify critical shortages of resources with up-to-date information Minimizes costly errors in estimating activity timing and resource requirements Promotes consistent level of detail in scheduling and resource allocation Reduces maintenance and storage costs through convenient location of information Allows coordinated, optimized scheduling and allocation of resources
Reporting of Costs and Accomplishments	Allows comprehensive summary of Asset Management activities via integrated data Helps track and manage costs and performance effectively using up-to-date data Improves the reliability of dollar values and production rates reported by the agency Promotes standardized cost and accomplishment reporting procedures Allows full attribution of costs and accomplishments to specific agency functions Facilitates rapid production of reports and graphs depicting costs and accomplishments Reduces storage and maintenance costs for large amounts of data
Performance Evaluation	Provides for the immediate feedback necessary to improve performance Gives higher level of confidence to the calculated performance measures and indicators Promotes consistent performance measures throughout the agency Reduces time to calculate and evaluate multiple performance measures for many assets Provides flexibility to conduct different types of analysis with the data Allows quick comparison of assets, resources, personnel, and activities

HOW TO INTEGRATE DATA

The data integration process can be extremely involved and challenging, especially for organizations that have a long history of stand-alone files or rarely share data across database systems. However, without fully integrated data, the objectives and benefits of comprehensive Asset Management will be difficult to realize.

A careful and thorough analysis of its Asset Management activities can help a transportation agency identify its needs, priorities, and capabilities in regard to data integration. Before embarking on the integration task, an agency may find it useful to form a data integration team consisting of database users, asset managers and decisionmakers, information technology and database management professionals, and other key stakeholders in the Asset Management process. The agency may also need to engage the services of an external group of data integration experts or consultants to assist in any or all stages of the process.

A generic framework for the key activities involved in data integration and the corresponding elements to consider for each activity are shown in Figure 2, The Data Integration Process. The process begins with a requirements analysis, which includes identification and analysis of the target data to be integrated, the Asset Management business processes that the integrated data will support, the requirements of the managers and users of the database, and other technical and organizational considerations, such as existing information systems infrastructure (hardware and software) and staffing capabilities. The next step, data and process flow modeling, uses the information obtained from the requirements analysis to build diagrams depicting the flow and use of data across the agency. Alternative data integration strategies can then be identified and evaluated on the basis of all the information assembled. A strategy can be chosen, and its detailed database design specifications and plans developed. The final step involves development, testing, and implementation of the integrated data strategy. More detailed descriptions of each activity shown in Figure 2 are provided below.

Approximately 60,000 miles of State-maintained highways on Virginia DOT's fence-to-fence right-of-way will be included in the Inventory and Condition Assessment System (ICAS). ICAS is a comprehensive information system that includes all maintainable highway assets. By virtue of its design, ICAS will support comprehensive Asset Management and program development.

REQUIREMENTS ANALYSIS

A requirements analysis is the first and most important stage of data integration. In this stage an agency will identify the requirements of the data integration system: the business processes that will be supported, the data that will be shared, the goals the agency is trying to achieve, and the constraints or challenges that are expected to impact the process. Depending upon the size and extent of integration, the analysis can be quite complex and time-consuming. Following are brief explanations of specific data integration requirements.

Business Processes

Typical business process functions that will be supported by integrated databases include inventory, data handling, and decision-support processes and systems for pavements, bridges, tunnels, roadway hardware, equipment, and other physical transportation assets (see Table 1 for a list of these processes). The requirements analysis stage of data integration characterizes each business process, the types of information it uses and produces, and the individuals involved. For example, if data integration were to support roadway sign inventory and condition assessment, the requirements analysis would identify the types of information associated with the inventory and condition of the signs, such as location, sign type, and reflectivity. The analysis would also identify the specific

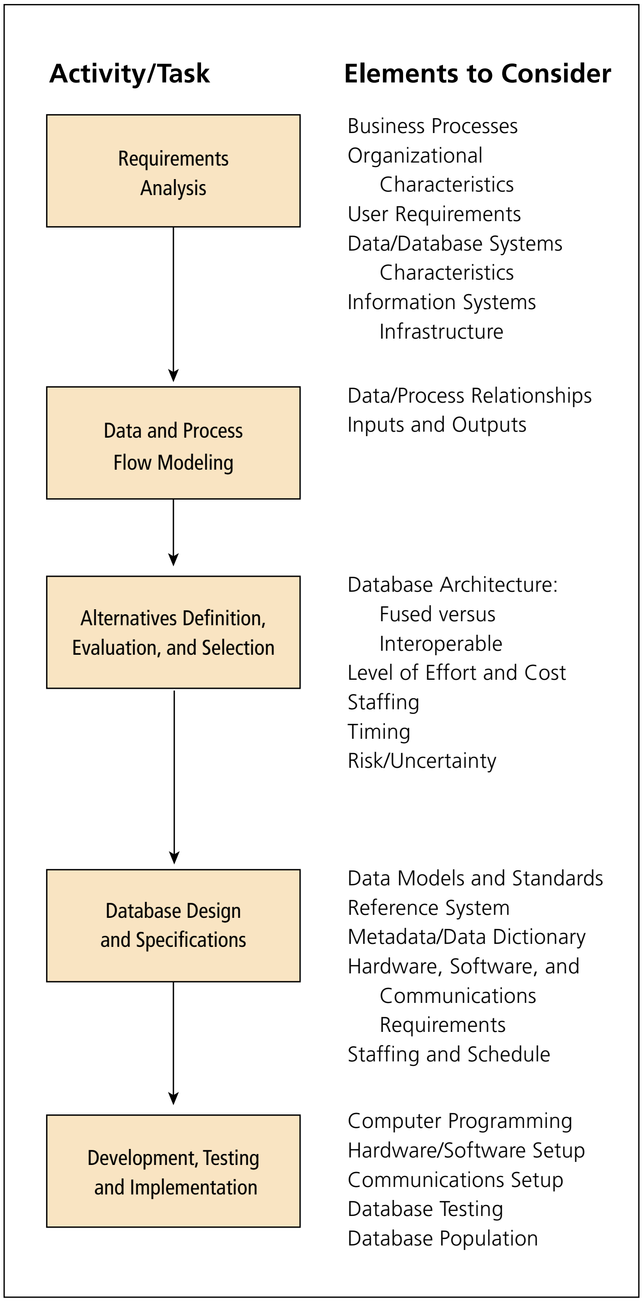

Activity/Task
Elements to Consider
Requirements Analysis
Business Processes
Organizational Characteristics
User Requirements
Data/Database Systems Characteristics
Information Systems Infrastructure
Data and Process Flow Modeling
Data/Process Relationships
Inputs and Outputs
Alternatives Definition, Evaluation, and Selection
Database Architecture: Fused versus Interoperable
Level of Effort and Cost
Staffing
Timing
Risk/Uncertainty
Database Design and Specifications
Data Models and Standards
Reference System
Metadata/Data Dictionary
Hardware, Software, and Communications Requirements
Staffing and Schedule
Development, Testing and Implementation
Computer Programming
Hardware/Software Setup
Communications Setup
Database Testing
Database Population

FIGURE 2. THE DATA INTEGRATION PROCESS

agency staff involved in the sign inventory and condition assessment process, including field crews, sign managers, and maintenance managers in the districts and headquarters.

Organizational Characteristics

Every transportation organization is unique in its structure as well as its management and operating characteristics. The requirements analysis should reflect the characteristics of the agency, particularly the various groups that will be impacted by data integration. Each group's business processes should be known, as well as the relationships within and among groups, the skills and capabilities of the staff, and the receptivity of the individuals to data integration—for example, whether they feel it is something they need to improve their effectiveness.

User Requirements

Ideally, all users of the database will be involved in the data integration process. Considering their inputs and requirements in the design process will allow the integration strategy selected to meet their data requirements. Some of the information that will be acquired from the data users includes where and how they obtain the data, the business processes and information system supported by the data, and any concerns they have about integrated databases. Adequate knowledge of the data users and incorporation of their requirements in the integration strategy will also help achieve cooperation and buy-in. Key to a successful data management strategy is that it includes information relevant to the requirements of all the data users in the agency.

Data and Database Management Characteristics

An essential component of the data integration requirements analysis is familiarity with the existing data and database systems within the agency. The analysis involves gathering all relevant information pertaining to the data:

- Where the data come from and who collects the data
- Method and frequency of collection
- Reference system or systems used
- Structure, format, and size of the data
- How the data are transmitted, processed, and stored
- General quality of the data in terms of accuracy, completeness, recency, and redundancy
- How the data are used (e.g., in which business processes)
- Applications that draw data from the databases (e.g., pavement management system, bridge management system)
- Types of reports produced

The characteristics of the various applications and information systems that will use the integrated data are identified and described below.

Michigan DOT started its data integration efforts by deciding to adopt a new location reference system for use by all transportation agencies in the State. The DOT conducted a detailed database design and migrated to a new database environment from the original mainframe systems, abandoning all existing data sources and related software applications.

Information Systems Infrastructure

Another component of requirements analysis is a review of the agency's existing information systems (IS) to determine the appropriate software, hardware, and communications strategies for integrating databases. This analysis will enable the agency to evaluate its readiness for data integration and, in particular, the suitability of the current IS infrastructure to each of the potential data integration strategies. Useful information might include an inventory of existing computer programming environments and database management or mapping software or servers, as well as computer hardware and operating systems. To avoid costly new purchases or acquisition of incompatible hardware, software, and other equipment, the agency may consider data integration alternatives that match or complement its existing IS technologies, thereby minimizing the cost and level of effort required.

DATA AND PROCESS FLOW MODELING

As shown in Figure 2, data and process flow modeling follows the requirements analysis phase of data integration. Specifically, the information obtained from the user requirements analysis can be used to develop diagrams depicting the flow of data within and among the business processes. The objective of data and process flow modeling is to create a picture of the relationships between the data and the business functions that the data support. Data flow diagrams help database engineers and analysts determine the design specifications for the integrated data. Flow diagrams consist of data, business processes, and relationships, which are represented by arrows. The direction of the arrows indicates where the information is going or which process is dependent on another.

The example in Figure 3, Section of Data/Process Flow Diagram for Culvert Management, depicts data and process flow for three interrelated culvert management functions. Note that the diagram shown is only one section of a larger flow diagram that would depict all the culvert management functions involved. In Figure 3, evaluation of maintenance requirements is dependent on both culvert inspection and condition assessment.

All data and business processes identified in the requirements analysis can be captured in flow diagrams, showing the movement of data within an organization. The diagrams are used by integration analysts to develop appropriate integration design plans. A variety of software exists to systematically create data and process flow diagrams.

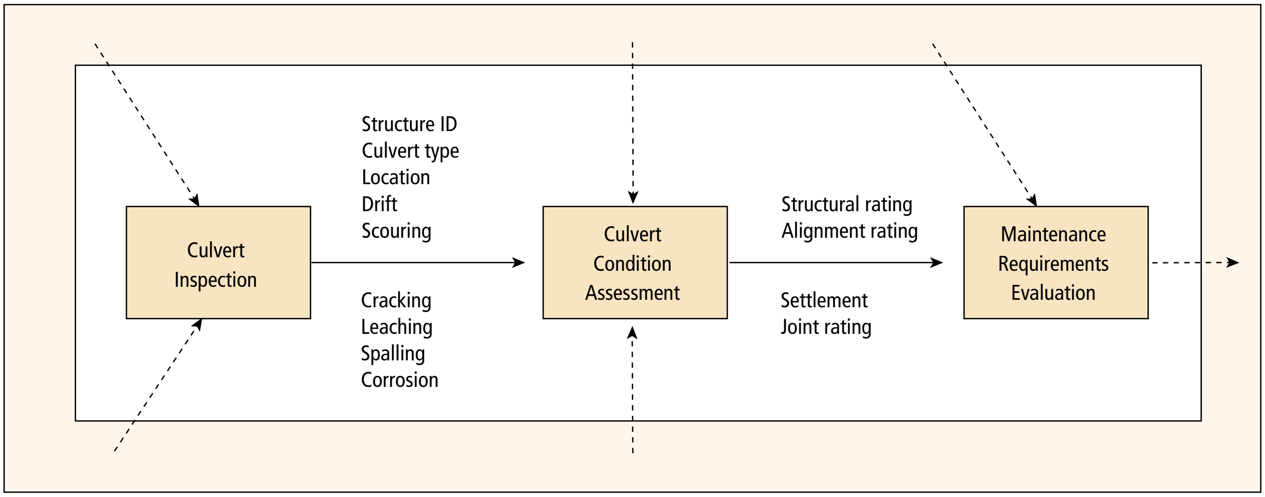


FIGURE 3. SECTION OF DATA AND PROCESS FLOW DIAGRAM FOR CULVERT MANAGEMENT

ALTERNATIVES DEFINITION, EVALUATION, AND SELECTION

The information obtained from the requirements analysis and flow diagrams provides the basis for the next step in data integration: identification of feasible integration alternatives. In general two alternative approaches exist: fused databases and interoperable databases. Data fusion—also referred to as data warehousing—pertains to one-time integration that combines information from multiple sources. These sources of fused data may be cast aside when data are migrated to the warehouse, or they may continue to exist to serve specific business processes. Alternatively, interoperable database systems—also referred to as federated or distributed systems—consist of databases that communicate among themselves via multidatabase query. Interoperable databases involve an alternative interface through which a data source (e.g., an existing database) can be viewed and manipulated. The two database options are depicted in Figure 4, Data Integration Alternatives, and discussed below.

In a fused data (or warehousing) environment, all data reside in a single database server with large processor and data storage capacity, and all personal computers and terminals connect to this server to access the data and perform any data processing function supported by the data warehouse. In the interoperable database environment, the data are distributed in computers or database servers located in various sites that are linked through a computer network. This network allows data from one computer to be accessed by or transferred to another computer.

Fused Databases

Data fusion, or warehousing, gathers selected data from various data sources, cleans the data by removing inconsistencies, and exports the data to a centralized database that is a replicate of data at the other locations. As illustrated in Figure 4, a data warehouse allows users access to vast stores of data through a collection of data views that are derived directly from an array of integrated databases. The data warehouse program uses a common user interface for the relevant subsets of all component databases that feed the warehouse and specifies the rules for doing data fusion.

Data fusion often requires conversion of a database and its applications to a new database environment, that is, from one data format to another. Data must then be shipped from the old (or legacy) system(s) to the new ones, often through data reengineering and other integration methods. The old systems may continue to be used after their data have been converted to the data warehouse, or they may be fully migrated, in which case they are abandoned when the warehouse is complete. A potentially difficult task involved in Asset Management data warehousing is handling data from legacy management systems. Often the data are in different databases or formats and reside in disparate sources and applications. Legacy systems generally do not have the means to ensure the integrity of data in the database, making it hard to map information from those systems to the new ones.

There is no single approach to building a data warehouse that will meet the requirements of every organization. To effectively support a large data warehouse, the database management system needs to handle the accumulation and management of large amounts of data, while still providing easy and rapid access to information. Since the technology is continuously changing, and as organizations are learning more and more about developing data warehouses, the likely approach to data warehousing will be an evolutionary one.

> Maine DOT is developing a geographically linked data warehouse called Transportation Information for Decision Enhancement (TIDE). TIDE will provide access to the agency's legacy databases, maintain historical data for trend analysis, provide both standard and ad hoc query environments, and have spatial query capabilities.

Interoperable Databases

An interoperable database system, also called a federated or distributed database system, is a collection of separate and possibly diverse interoperating database systems over multiple sites that are connected through a computer communications network. The objective of this approach is to create an integrated model so all linked databases that are members of the federation appear to form a single database—the federated view. Interoperable databases are characterized by transparent access; that is, users can access the database without having to learn the data model or write transactions in the language supported by the database. For example, a user of one database format in the agency's finance department may access a database of a different format in the engineering department. Likewise, database owners can share their databases with others without compromising database integrity. The autonomy of the owner's database is upheld despite the multiple accesses and manipulations of other users. Interoperable databases support different data models and execute transactions written in various data languages. The two key factors in successful implementation of an interoperable database are integrated data standards and distributed processing capabilities.

Creation of the federated view involves determining the data interface and linking individual databases to that interface. The integrated database format hides the complexity and distribution of the underlying component databases. This greatly simplifies the task of developing a database query and promotes a common understanding

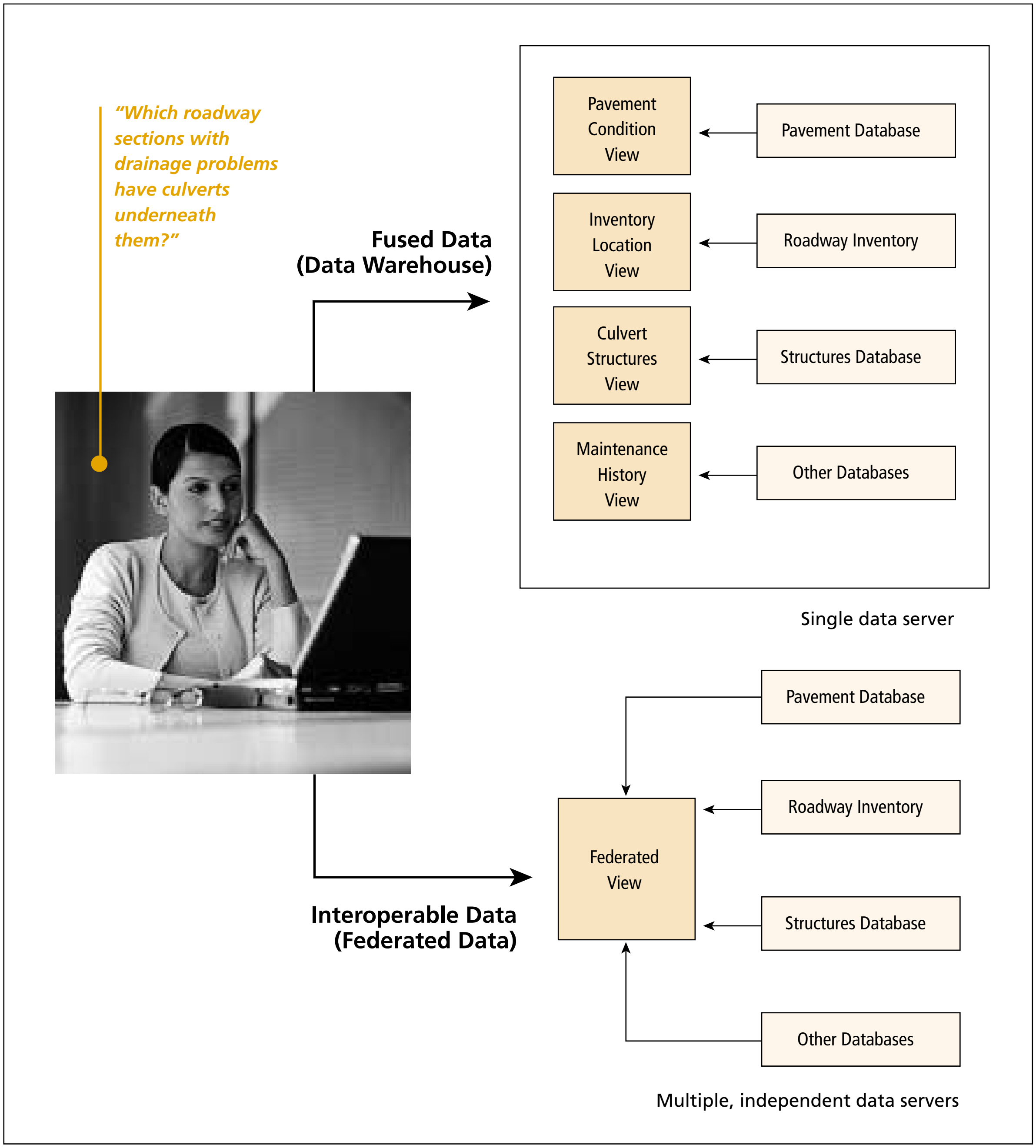


FIGURE 4. DATA INTEGRATION ALTERNATIVES

of the data. As shown in Figure 4, a user can make a query without regard to the location of particular data or the diverse representations used by the various databases.

Decentralized transportation organizations are ideal environments for interoperating database systems because they fit more naturally with the structure of such organizations. Database management approaches may be similar within a particular unit or division but are often distinct across divisions. The federated approach is appropriate when several databases exist (e.g., pavement database, bridge database) and there is a need to run agency-wide applications such as a maintenance management system. Its advantages over centralized (fused) database systems are that it provides more access to computer resources on the network, improves database availability, and allows data sharing while maintaining some measure of local database control. Federation can be used to provide access and preserve the investments in legacy systems of a transportation organization.

The disadvantages of federated databases include the difficulty of maintaining an integrated global model when thousands of databases are involved. Installation and configuration of federated databases also require much expertise. The databases themselves can be problematic due to differences in query dialects, functions supported, data types, and database versions. Moreover, federation and its associated communications systems require considerable tuning to maintain acceptable performance.

Evaluation and Selection of Alternatives

Table 2 summarizes the major advantages and disadvantages of fused and interoperable databases. Criteria that might be useful in evaluating the integration alternatives include required level of effort, time requirements, estimated costs including risks, and anticipated agency improvements or benefits. Other factors may be considered depending upon the needs of the agency and the constraints identified in the requirements analysis. Following acceptance of a specific approach, the next stage in data integration is the development of an implementation plan that includes a schedule of database design, development, and testing, as well as plans for software and hardware purchases.

TABLE 2. COMPARISON OF FUSED AND INTEROPERABLE DATABASES

Characteristic	Fused Database (Data Warehouse)	Interoperable Database
Number of Data Servers	One (central)	Multiple (distributed)
Location of Data Server(s)	Single site	Multiple sites
Data Replication	Yes	No
Advantages	Easy to manage and control the databases. Maximum data processing power (quick access to the database). Able to handle large amounts of data and processing requests. Provides data security.	Can keep data in independent locations and file servers (autonomy of sites). No reliance on a single site that can become a point of failure. Changes made to data at one location can propagate quickly to become visible at other locations. Unified description of all data—no need to know database models. Allows access to resources in the computer network.
Disadvantages	Requires considerable time and resources to implement. Data is generally in read-only format and cannot be updated online. Storage requirements can become a major problem.	Hard to support and maintain integrated (global) data model. Need to rebuild the database system every time data export protocols change. Requires rigorous procedures for database access and updates.

DATABASE DESIGN AND SPECIFICATIONS

The database design and specifications can be used to generate the detailed plans and methods in implementing the selected data integration strategy. They also produce the overall approach to the database development effort. Whether a fused or interoperable database environment is selected, these elements are included in the integrated database design: data models, standards, and reference systems; metadata and data dictionary; computer communication requirements, software, hardware, staffing, and data management requirements.

Data Models

The structure and configuration of the database are represented by the data model. Database design involves identifying appropriate data structures for either the fused or linked databases. Examples of data models include flat file, hierarchical, network, relational, and object-oriented models (see sidebar). The data and process flow diagrams developed earlier in the integration process will assist in selecting appropriate models for the asset management data. Existing models adopted by the agency for some of its data or database applications might be good candidates as integrated data models. Ideally, the data models selected will be directly compatible with applications that use the data. A data conversion or transformation routine can also be developed that would allow the applications to use the data in a different format.

Data Standards

Incorporating data standards in the database design and specifications helps facilitate the integration process by identifying or establishing acceptable rules for representing, accessing, manipulating, transferring, and reporting data. Some of these standards or rules may already exist in the agency and would be identified in the requirements analysis stage of the data integration process. The development of new standards requires evaluating both the data elements and the processes that draw information for specific applications. The three most common data standards are data representation standards pertaining to how a specific data would be stored in the databases (e.g., content, format); data access and manipulation standards including conventions for requesting information from the database (e.g., database communications protocol and database query language); and data transfer and reporting standards that specify the format that will be used to export the data from the database to an external destination such as an application or another database.

DESCRIPTION OF DATA MODELS

Flat File Model—A file structure involving data records that have no structured interrelationship. A flat file takes up less computer space than a structured file but requires the database application to know how the data are organized within the file.

Hierarchical Model—A data model that links records together like a family tree, but each record type has only one owner (e.g., a purchase order is owned by only one customer). Hierarchical data structures were widely used in the first mainframe database management systems. However, due to their restrictions, they often cannot be used to relate structures that exist in the real world.

Network Model—A special case of the hierarchical data model in which each record type can have multiple owners (e.g., purchase orders are owned by both customers and products).

Relational Model—Data items are organized as a set of formally described tables from which data can be accessed or reassembled in many ways without having to reorganize the database tables. Each table (sometimes called a relation) contains one or more data categories in columns. Each row contains a unique instance of data for the categories defined by the columns.

Object-Oriented Model—Defines a data object as containing code (sequences of computer instructions) and data (information that the instructions operate on). Traditionally, code and data have been kept apart. In an object-oriented data model, the code and data are merged into a single indivisible thing—an object.

Data Reference Systems

Reference systems are integral components of the database design as they are used by the database management system to link and relate separate data files. For highway assets the most common reference system used is the location reference system, particularly in terms of route number and milepost. Other location systems may be used, such as the coordinate reference system involving latitude and longitude. The location reference system allows not only the integration of spatially referenced data but also the mapping and analysis of information using geographic information systems (GIS) software. In addition to location reference, some data may need to be referenced by other coding systems, such as those used for accounting or inventory purposes.

> The most difficult and time-consuming processes in the development of Mississippi DOT's Transportation Management Information System were the establishment of an agency-wide common location referencing system and the subsequent conversion of data to conform to that system.

Metadata and Data Dictionary

In addition to defining the data models, standards, and reference systems, a key component of database design is the creation of metadata and data dictionaries, which are basically detailed descriptions of the data. "Metadata" pertains to data about the data. One aspect of metadata includes the data's meaning: what it represents in the real world as well as its formal names, definitions, integrity, and accuracy. The other aspect of metadata describes the physical nature of data: how it is stored, the data types (e.g., text, image, numeric), structure (e.g., relational, object-oriented), location, systems accessing it, and any other information to help the database analyst manage the data. The data dictionary is a subset of the metadata containing an organized catalog of the data files about the definition, type, structure, and other properties of the data such as the number of records or fields. The data dictionary facilitates the integration of data by ensuring a consistent definition and correct usage of data in the databases and by making clear distinctions among various data items.

Computer Communication Requirements

The process by which integrated data will be accessed by various users or clients from their computer terminals and workstations is also included in the database design and specifications. Access to the data by end users and application programs is normally carried out through a computer interface. Depending on whether a fused or interoperable database is selected, the user may access data directly from a warehouse in which the database resides or dial up to a computer network to access the data from other computers or database servers and use these data on a local machine. A computer network is a system where multiple computers share the resources of all computers connected to the network via a high-speed data communications link. The computer network allows the databases stored in one location to be accessed by various users working on separate computers in different places. These users can communicate with each other or exchange information via the network. The communications requirements for integrated data that will be identified in the database design include dial-up and communications procedures and other components, such as software, needed by the computer network.

> The Turnpike District of the Florida DOT is developing Turnpike Asset Management System, a Web-based, graphically driven, fully integrated asset management system that can be accessed by District personnel through the District intranet.

Software and Hardware Requirements

Software and hardware requirements for the integrated database depend upon the database design and specifications described above. These include software and hardware choices and requirements for database servers, network communications, data mapping, user interfaces, computer operating systems, and programming environments. Agencies have the option of building the database management system from scratch or adopting a commercial software package, which may be customized.

Database Server. The choice of the server software and hardware is important because they are used to store and manipulate the databases. Key factors to consider are the maximum number of users expected to access the database at any one time, the level of uptime needed, the types of programs that will be accessing information from the database, the hardware and operating system the server will be using, and the level of in-house expertise the organization already has in a particular server environment. The server ideally would have adequate speed and storage capacity to handle large and complex data processing jobs.

Recent trends in the software market are making it easier and less costly to implement integrated databases. Many software packages are available that will support hundreds or thousands of simultaneous users logging on from various locations. They all have sophisticated security controls and can be customized to fit particular business needs.

Server hardware is also becoming cheaper and faster. While organizations might want to buy multiple servers for redundancy, there is not much need to make significant investments in hardware to obtain improved performance. The Internet also has substantially influenced the database market. Database vendors are adding Internet support and seamless Web connectivity to their products. Most vendors have incorporated Internet-enabling utilities to some extent to make their products more programmable and compatible with Web-based applications. Data integration, which was once so expensive and obscure that only the largest organizations could afford to use it, has really come of age.

Mapping Software. Spatial reference and mapping software, which is used to display and analyze location data, is collectively referred to as GIS software. The spatial nature of most transportation data makes GIS a powerful tool for Asset Management. GIS software is used for constructing spatial databases of transportation networks and features, conducting various types of analyses and applications on the spatial data, and integrating many management and decision-making information and processes. Some GIS products have external database integrators that enable them to coexist and be integrated with an organization's IS infrastructure. This functionality provides GIS users with the ability to access and use data from a number of relational database management systems. Existing transportation spatial databases or warehouses developed by a number of highway agencies use GIS software and modules to link databases or perform specific database functions such as querying or reporting. GIS software options include several commercial products or suites of products. Each software product offers various data management, analytical, and reporting capabilities. Some products are designed for Web-based mapping and analysis applications. The software runs in different computer operating systems and network environments.

Commercial Off-the-Shelf (COTS) Packages. In lieu of building Asset Management applications and data integration routines or software from scratch, agencies have the option to acquire prepackaged COTS software that has the basic, generic functions of Asset Management processes and can be tailored to specific agency applications. This software ranges from large, enterprise-wide suites of applications, commonly referred to as enterprise resource planning (ERP) software, to other products that can be used for several asset types or several Asset Management processes. COTS data integration software may save time and money when compared with writing customized programs.

The Virginia DOT adopted multimodule software to manage its Inventory and Condition Assessment System database and to provide the core functionality for all system users.

An ERP system is a multimodule application software used to support a broad set of activities to help an agency manage the important parts of its business. ERP software solutions have been growing in popularity over the past few years. Typically the software operates on the client-server structure, but applications also operate in conjunction with midrange and mainframe computers. ERP systems are typically Windows-based and are grouped into various functional modules such as accounting, finance, manufacturing, human resources. However, these modules work together to control and analyze data. They have the ability to organize the agency's information and analyze it in real time. This is a tremendous benefit to data management. A variety of COTS software packages are available that are focused on transportation data management and integration.

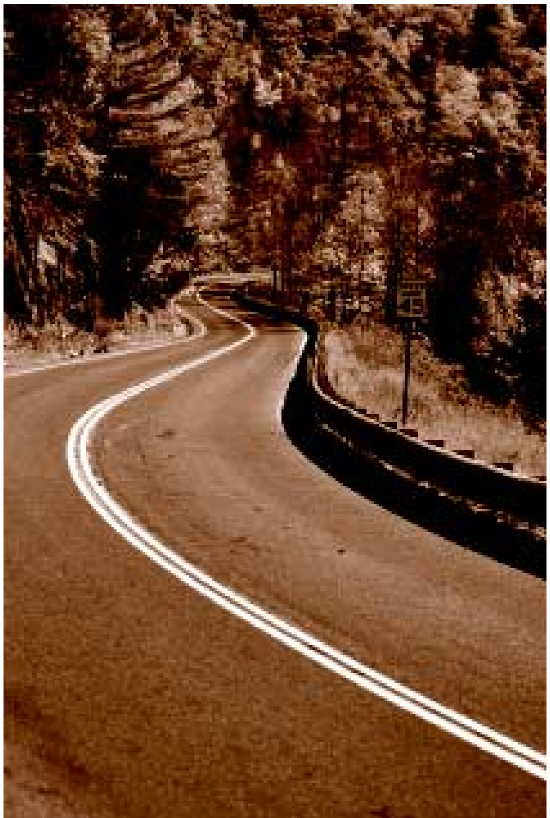
Staffing and Data Management Requirements

Finally, during the database design and specifications stage of data integration, management and administrative responsibilities for the integrated databases are established by identifying the people who will be managing database programming, prototyping, and testing (database development), software and hardware purchases (procurement), computer network setup (systems administration), and database management, maintenance, and upkeep.

DEVELOPMENT, TESTING, AND IMPLEMENTATION

The last stage in data integration is software development and system implementation. These include prototyping and use case applications development, computer systems and network communications setup, and populating the database with data. Development activities include testing, evaluation, and modification of database models, data management applications, and communications interfaces. It is advisable that the development approach be as modular and incremental as possible in order to accommodate future additions or changes to the database or any component of the integrated environment.

Database developers usually create prototypes and use case applications to implement the integrated database models. Prototyping and use case modeling ideally are performed in tandem in order to focus on the data users and their use of the system. The initial version of the prototype will consist of major program modules written to move data back and forth between the screens, the database, reports, and the inputs and outputs used to communicate with other data systems. At first, these prototypes may do little data processing. As the prototyping continues, newer versions of programs that perform full-blown data processing will replace the original versions.



Prototype development for Maine DOT's Transportation Information for Decision Enhancement data warehouse worked well to identify potential problems early in the process and provide better understanding of outstanding issues requiring future resolution.

The computer network communications components of the integrated system are put in place during the programming and software development stage when the interface or connections between databases are ready for setup and testing. Again, depending upon the scope of integration and type of network configuration chosen, implementing the network communications may require significant effort and agency resources to implement. Specifically, the more complicated the communications requirements of the integrated database system are in terms of computer connectivity, data access rates, data retrieval and processing, or system flexibility and reliability, the more time and money will be spent.

The final step in the process involves filling in (populating) the database with the necessary data, both historical and newly acquired.

SUMMARY

Transportation agencies may use the methodologies presented here to successfully undertake a data integration initiative. The procedures described for requirements analysis, data and process flow modeling, alternatives evaluation, detailed database design, software development, and implementation can help support an agency's data integration requirements and business processes. Additionally, the following four strategies may serve as general guidance for most situations:

1. Use a data environment that facilitates making changes in database functions or adding new data sets.
2. Adopt an incremental development approach to ensure flexibility for change and provide sufficient time to test and upgrade the integrated databases.
3. Involve the database users during all stages of design and development to benefit from their input and assure cooperation and buy-in.
4. Select hardware and software that meet the goals of the database management system, data users, and agency.

CHALLENGES TO DATA INTEGRATION

Despite the challenges presented by the typical situation, where many disparate databases are present, data integration efforts often begin with the goal of perfect coordination of database servers and clients. Those efforts are likely to fail or end prematurely if the full scope of complexity is not recognized in the requirements analysis stage of the process. The most common technical and organizational impediments encountered when trying to integrate enterprise data are described below with some recommendations for overcoming these hurdles.

HETEROGENEOUS DATA

Most transportation agencies have large quantities of variable, heterogeneous data. Data heterogeneity usually results from the presence of internal legacy systems that use different data formats. Many legacy systems were constructed around flat file, network, or hierarchical databases (see sidebar, page 19). Newer generations of database systems utilize relational data. Data in different formats from external systems continue to be added to legacy databases to enhance the value of the information. Each of these generations, products, and home-grown systems has unique properties with regard to data storage and data extraction. Thus, data integration often involves coping with heterogeneity in data, and in some cases the whole effort can become a major exercise in data homogenization.

To mitigate this problem, a thorough analysis of the characteristics and uses of data is necessary. The model chosen to represent the data in either a federated view or a data warehouse environment will be based on the requirements of the business process applications and other uses of the data. If data are to be transformed to a new format, the database developer sees to it that the applications can use data in this format or that standard procedures are in place to convert the data to another format supported by the applications. The effort required to tie disparate data together in an interoperating database system or to migrate and fuse highly incompatible databases into a standard model can be tremendously painstaking and can overwhelm the integration task. However, advances in software technology are helping to minimize the problem with a series of data access routines that allow structured query languages to access nearly all database management and data file systems—relational or nonrelational. These processes make it possible for data access tools to read or acquire data stored in database management systems.

BAD DATA

Data quality is a primary concern in data integration initiatives. Those who avoid cleaning up legacy data prior to conversion and integration can face serious data problems. Legacy data impurities have a compounding effect since, by nature, they tend to concentrate around the biggest data users. Even a small percentage of legacy data corruption can invalidate the information and the decisions that are drawn from it. In organizations with complex data processing systems, it is not unusual for previously undiscovered data quality problems to emerge as soon as efforts to clean the data are initiated. Many find it necessary to install procedures to regularly audit data quality. However, in most agencies it is unclear who has responsibility for this task.

Since data quality is a continuous requirement of the data integration development and implementation efforts, it is best to establish it up front, at the beginning of the task, and assure that it is an ongoing responsibility. It is a good investment of time in the initial stages of a data integration project for the developers and users to jointly determine what quality checks will be made during the development phase and what quality checks and procedures will be needed on an ongoing basis.

LACK OF STORAGE CAPACITY

One of the most common data integration problems, particularly in regard to data warehousing, is the unexpected need for additional performance and capacity. This translates into two storage-related requirements—extensibility and scalability. Recognizing that disk requirements for data can be very expensive, many data integration architects are concerned that a data warehouse could easily grow to 20 times its size in a year and cause a major storage problem. With storage management costs at eight times that of initial hardware, the total cost can easily tip the cost/benefit balance for an integrated database. Adding massive quantities of data into the database equation likewise increases the chances of failure and could push the limits of the database software and hardware, forcing developers into using costly development techniques if a massive processing architecture is needed.

Many transportation agencies are starting to look at other ways to meet their large data integration storage needs. Alternative storage is becoming mandatory for warehouses that will grow in size. These alternatives hold the promise for making growing databases affordable. The disk drive market continues to offer lower cost per megabyte of storage space. High-performance storage disks, the foundations of typical data warehousing environments, may likewise follow the downward pricing trends of the overall storage market.

> One important consideration by the Michigan DOT in developing its Transportation Management System (TMS) was to allow for database expansion, recognizing that it would be impossible to anticipate every single data item that will be used for their business processes in the future.

UNANTICIPATED COSTS

The main costs associated with data integration include labor costs during initial planning and evaluation, programming, software, hardware, data acquisition, and storage and maintenance. An accurate and realistic cost estimate is needed to ensure that actual expenditures do not exceed the budget, or the project may suffer an unexpected demise. Unrealistic estimates can be caused by an overly optimistic budget or lack of experience in estimating these costs. Performance or capacity problems resulting from more users, more queries, or more complex database requirements than anticipated may require more hardware or effort than originally planned. Due to limited resources the project scope and time allotment may have to be extended without a change in the budget. In some cases, unanticipated and expensive consulting help may be needed. Moreover, extenuating circumstances such as delays caused by hardware or software problems, lack of staff, change in the business processes, and other factors may result in additional expenses.

The proportion of time that must be spent in extracting, cleaning, loading, and maintaining data can be significant as explained above for heterogeneous and bad data. Additional labor costs associated with these activities can become exceedingly out of proportion with the total costs and render the data integration project grossly underfunded.

Data integration analysts need to have a reasonably farsighted yet realistic approach for estimating project costs. Factors that have even the slightest likelihood of affecting the effort but have substantial consequences on the total cost should be taken into account when building cost projections. The objective is to anticipate and minimize any overruns, which may occur at every step of the way.

INADEQUATE COOPERATION FROM STAFF

Users may decide not to use an integrated database if it is not responsive to their preferences. Likewise, many user groups within the agency may have developed databases on their own that meet most of their key reporting needs. These systems often will have been built independently of the agency's IS group. When the new integrated databases begin to subsume the functions of the stovepipe (standalone) databases, the owners and users may be skeptical about whether the organization can do as good a job supporting the user's reporting needs as the original owners of the data did on their own. If the goal of data integration is to electronically link and automate the production of spreadsheets that a user has been manually creating, the user could feel threatened by the prospect of automation. Privacy and confidentiality may be another cooperation hurdle. That is, one business unit may not want another unit to see its data. Another potential impediment arises when division personnel do not want headquarters personnel to see detailed division data, concerned that they may lose control over what they feel is their own data.

Lack of financial, personnel, and consultant support from high level management will easily defer any effort to integrate data. Top management not only has to approve the strategic plan and resources associated with the project, but also has to help validate and communicate the need for data integration to everyone in the agency. Any large-scale data integration project, whether a data warehouse or an interoperable database system, needs executive management on board. Their involvement can break down the barriers in organizations. Any project lacking executive management participation has a high probability of failure.

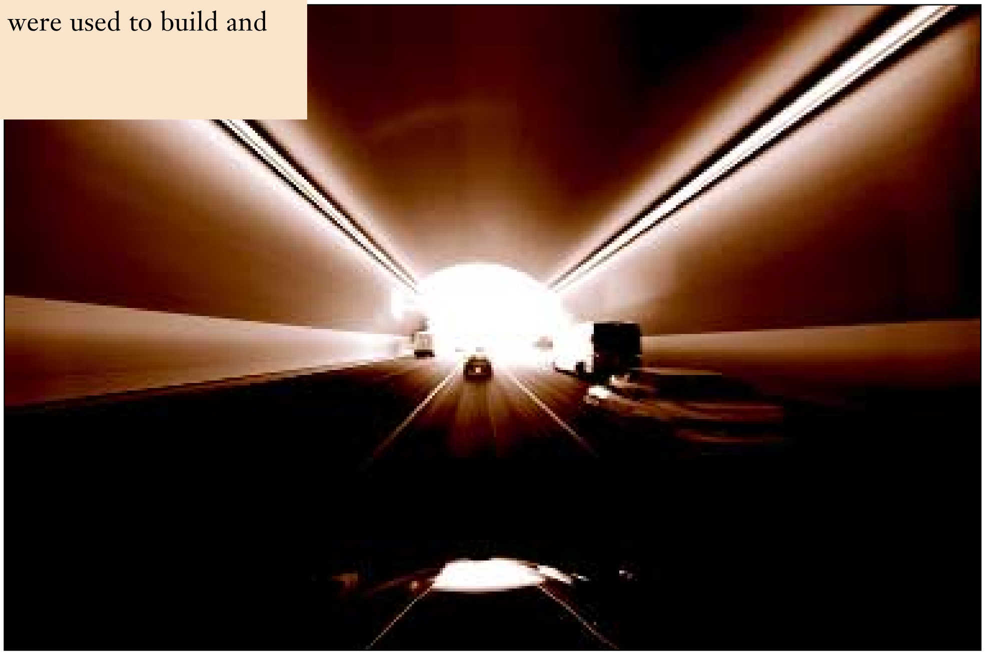
Informing, consulting, and involving the different players during the requirements analysis stage as well as subsequent stages of the integration process can help to ensure full support and cooperation. It can be overwhelming at times to address each user's concerns when designing an integrated database, but in the end this will influence the success or failure of the project. Taking it a step further, users need to have a personal stake in the success of the project and ownership in the final product. It is also important to educate the users on the fundamentals and processes of data integration—to teach them its benefits as well as its limitations so they can properly gauge their own expectations.

Florida DOT's Turnpike Asset Management System (TAMS) crosses all departmental lines within the Turnpike District. Building TAMS required buy-in from the top down by all affected departments and then a certain degree of autonomy in the implementation phase. User focus groups, which consisted of personnel from the staff level to the management level, were used to build and ensure consensus.

LACK OF DATA MANAGEMENT EXPERTISE

Integration can be impeded when sufficient knowledgeable staff and data management experts are not available. An integrated data system requires a considerable amount of time to fully develop, and it takes a while to gain experience with the common problems that develop at different stages of a data integration project. A knowledgeable and experienced data integration project leader and an expert data manager are needed to design a modular, robust, and maintainable architecture that can accommodate expanding and changing transportation decision-support requirements. At present, there are a limited number of people within transportation organizations who have experience and expertise working with the full life cycle of data integration projects.

Transferring historical data from an existing independent file to an integrated data system can require the knowledge of the original personnel who may have long since left; the turnover rate for these positions tends to be very high. Despite the best efforts to design a system that minimizes maintenance demands, many systems require a great deal of maintenance once they are in production. It is important to realize and accept that the more successful data integration is with users, the more maintenance it may require. Thus, adequately trained personnel are needed to support the care and feeding of the system.



CONCLUSION

When carefully planned and properly applied, data integration can bring substantial benefits to any transportation agency. This Primer provides information to assist agencies in undertaking a data integration initiative that will support their business processes, including an Asset Management decision-making framework. However, there are a variety of hurdles that can potentially cause a data integration effort to fail. Recognizing these hurdles and knowing how to deal with them will help minimize the risk of failure. Managing the cost and level of effort associated with data integration calls for considerable knowledge and foresight in the planning and analysis stages of the project. Database users are key players during all of the stages. Data integration is a continuous and evolving process that has to respond to the changing needs, business processes, and operating environments of the agency.

For further information and additional copies of this document, contact:

Office of Asset Management
Federal Highway Administraion
U.S. Department of Transportation
400 7th Street, S.W., HIAM-1
Washington, DC 20590

Telephone: 202-366-9242
Fax: 202-366-9981

U.S. Department
of Transportation